**Titanic Dataset**

**1. Introduction and Problem Definition**

The Titanic disaster of 1912 is one of the most infamous tragedies in history. The sinking of the ship resulted in the loss of over 1,500 lives, making it a compelling dataset for exploring predictive modeling. The Titanic dataset, which contains demographic and other relevant information on passengers, is a popular dataset for beginners and experts alike in machine learning. In this article, we will walk through the steps required to predict survival outcomes using this dataset.

**Problem Statement**

The goal of the Titanic prediction problem is to predict whether a passenger survived the Titanic disaster based on various features such as age, sex, class, and fare. This problem is a **binary classification** task, where the target variable is "Survived" (1 if the passenger survived, 0 if not).

**Problem Statement**:

* We can predict who survived the Titanic disaster
* Input data: Features such as name, age, gender, passenger class (Pclass), etc.
* Output data: Binary classification of survival (1 or 0).

**Evaluation-Metric**:  
We will evaluate model performance using **accuracy**, **precision**, **recall**, and **F1-score** to measure how well the model predicts survival outcomes.

**2. Data Analysis :**

Before jumping into building models, it's essential to understand the data and its structure. The dataset contains several features, some of which may be more useful than others for prediction. Let's take a look at the key features:

1. **PassengerId**: Unique identifier for each passenger.
2. **Survived**: 0 = No, 1 = Yes (This is the target variable).
3. **Pclass**: Passenger class (1st, 2nd, or 3rd).
4. **Name**: Passenger's name.
5. **Sex**: Gender (male/female).
6. **Age**: Passenger's age.
7. **SibSp**: Number of siblings or spouses aboard the Titanic.
8. **Parch**: Number of parents or children aboard.
9. **Ticket**: Ticket number.
10. **Fare**: Fare paid by the passenger.
11. **Cabin**: Cabin number.
12. **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

**Key Insights from Initial Data Analysis:**

* **Missing Data**: Age, Cabin, and Embarked have missing values, which need to be handled.
* **Categorical Variables**: Features like 'Sex', 'Embarked', and 'Pclass' will need encoding before being used in machine learning models.
* **Imbalanced Target**: The dataset is imbalanced, with fewer survivors than non-survivors. This can affect model performance and evaluation.

Below is a summary of the key features:

* **Pclass**: Indicates the passenger’s socio-economic status, with 1 representing the upper class, 2 the middle class, and 3 the lower class. This feature is likely a strong predictor of survival, as wealthier passengers may have had better access to lifeboats.
* **Sex**: This is a categorical feature that needs encoding into a numeric format. Preliminary analysis suggests that females had a much higher survival rate than males.
* **Age**: Age is a continuous feature that represents the age of each passenger. It contains missing values, so proper imputation strategies must be applied.
* **SibSp and Parch**: These features represent family size aboard the Titanic, with SibSp indicating the number of siblings and spouses, and Parch the number of parents and children. These features can be combined to create a "FamilySize" feature.
* **Fare**: Fare represents the ticket price paid by the passenger. This continuous feature could potentially be related to survival, as passengers who paid higher fares may have had better accommodations and lifeboat access.
* **Embarked**: The port of embarkation is a categorical feature with three unique values: Cherbourg, Queenstown, and Southampton. This feature may indicate geographical or socio-economic distinctions between passengers.

**Initial Data Exploration**

**Missing Values**

The dataset contains missing values in several important features, such as **Age**, **Cabin**, and **Embarked**. Handling missing data is crucial before feeding the data into machine learning models.

* **Age**: Approximately 20% of the Age values are missing. Missing values can be imputed using strategies like mean, median, or more sophisticated methods such as imputation based on similar records (e.g., passengers with similar Pclass and Sex).
* **Cabin**: The Cabin feature contains a significant amount of missing data (about 77%). Given the sparsity of the data, it may be better to drop this feature or use a binary flag indicating whether a cabin number is known.
* **Embarked**: A few missing values in the Embarked feature can be filled using the mode, as it's a categorical variable.

**3. Exploratory Data Analysis(EDA) :**

Exploratory Data Analysis helps us uncover trends, patterns, and relationships between the features and the target variable. Some common EDA techniques include visualizations such as histograms, box plots, heatmaps, and pair plots.

**Key Insights from EDA:**

1. Gender: Females had a significantly higher survival rate compared to males.
2. Pclass: Passengers in 1st class had a higher survival rate than those in 2nd or 3rd class.
3. Age: Younger passengers (especially children) were more likely to survive.
4. Fare: Passengers who paid higher fares tended to survive at higher rates.
5. Embarked: Those who embarked at Cherbourg (C) had a higher survival rate compared to the other ports.

**EDA Concluding Remarks:**

* Sex and Pclass are strong predictors of survival.
* Fare and Age also seem to have a significant impact on survival, although further analysis is required to handle missing age values.
* SibSp and Parch show some relationship with survival, but their influence is weaker compared to other features.

**4. Pre-processing pipeline :**

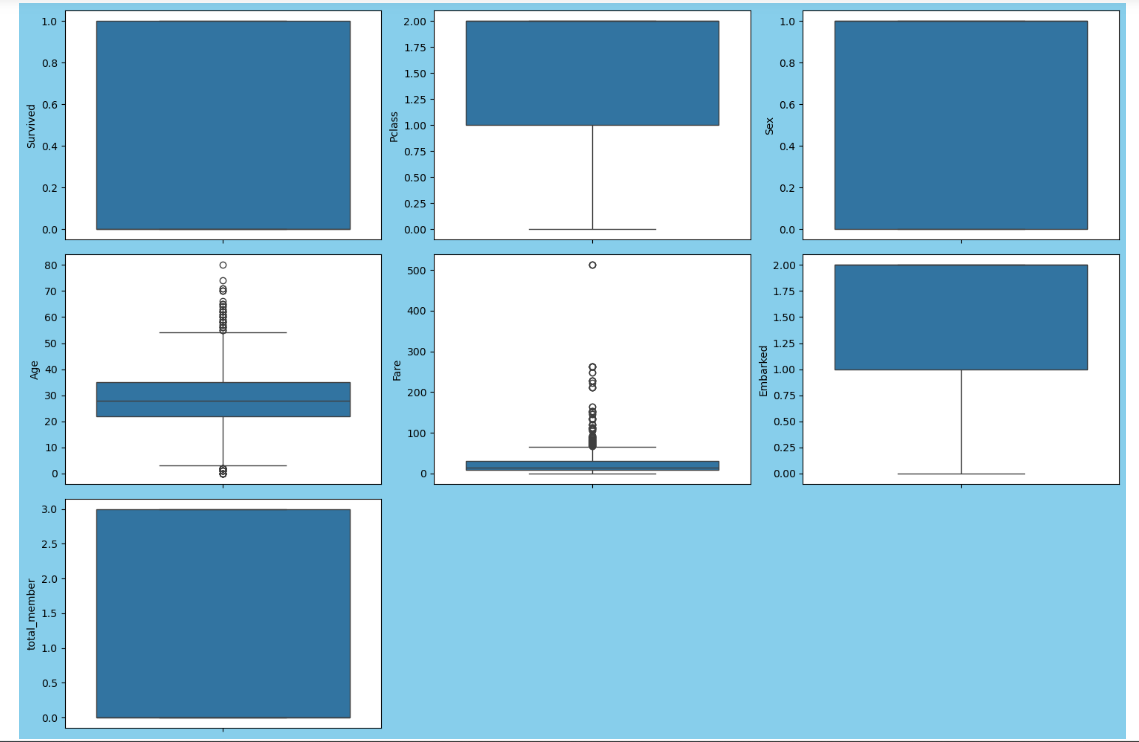
Pre-processing is a crucial step in preparing data for machine learning models. The Titanic dataset requires handling missing data, encoding categorical variables, and scaling numerical features. The following steps summarize the pre-processing pipeline:

1. **Handling Missing Data**:
   * **Age**: We can use either the median or an age-specific imputation (e.g., filling missing age values based on passenger class and gender).
   * **Embarked**: The missing 'Embarked' values can be filled using the mode (most common port).
   * **Cabin**: Since most of the 'Cabin' data is missing, we can either drop the feature or create a new feature indicating whether or not the cabin information is available (binary flag).
2. **Encoding Categorical Variables**:
   * **Sex**: Convert 'Sex' to numerical values (0 for male, 1 for female).
   * **Embarked**: Use one-hot encoding for the 'Embarked' feature, creating three new binary columns.
3. **Feature Engineering**:
   * **Family Size**: Combine 'SibSp' and 'Parch' to create a new 'FamilySize' feature, which may help capture the relationship between family size and survival.
   * **Title Extraction**: Extract titles (e.g., Mr., Mrs., Miss) from passenger names, as titles can be related to social status and survival chances.
4. **Scaling Features**:
   * Features like **Fare** can have wide ranges, so scaling using techniques like **StandardScaler** or **MinMaxScaler** can improve model performance.
5. Handing Imbalance Dataset :

Loan datasets often exhibit an imbalance between servived and non servived. This imbalance can skew model performance, causing the model to predict the majority class (loan approvals) more frequently.So we can now apply **SMOTE (Synthetic Minority Over-sampling Technique)**. This method oversamples the minority class by creating synthetic examples based on the existing data.

5. Checking the Outlier and Skewness :

Outliers are the abnormal data points that are significantly different from actual data points that form a skewness.Now observe the continuous variabled feature to detect the outlier.If any outliers present then remove that outliers for stable model performance and also remove the skewness of any related data points that are formed normal distribution.



6. Multicolinearity check :

Multicolinearity occurswhen two or more independent variables have high corelations with one another,it affect the bad impact of model performance.So remove the multicolinearity of independent features.

**5. Building Machine Learning Model :**

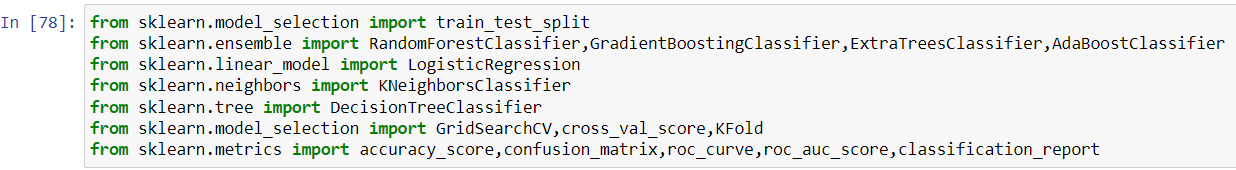
Now that the data is clean and preprocessed, we can build machine learning models. Several classification algorithms can be applied to predict servival rate.List of the algorithms are used to predict the servival rate :

1. **Logistic Regression**:
   * A simple yet powerful classification model that works well when the relationship between the features and the target is mostly linear.
2. **Decision Tree**:
   * A tree-based model that splits the data based on feature values. Each split creates branches, eventually leading to a classification decision at the leaf nodes.
3. **Random Forest**:
   * Random Forest is an ensemble method that builds multiple decision trees using different subsets of data and averages their results to improve accuracy and reduce overfitting.
4. **Gradient Boosting**:
   * Gradient Boosting models build sequential trees, with each new tree correcting the errors of the previous ones.

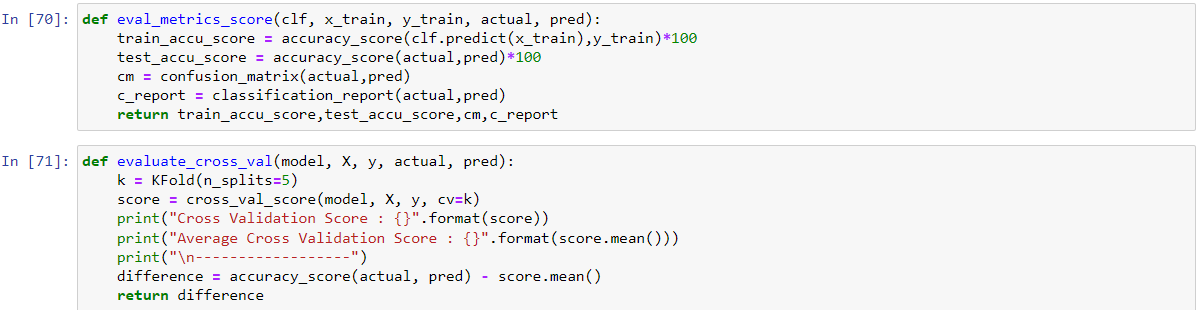
**Model Evaluation Metrics:**

For this classification task we check which model is best based on below metrics,

* **Accuracy**: Measures the proportion of correct predictions.
* **Precision**: The ratio of correctly predicted positive observations to the total predicted positives.
* **Recall (Sensitivity)**: The ability of the model to capture all actual positives.
* **F1 Score**: The harmonic mean of precision and recall, useful when classes are imbalanced.



**Now we check which model is perform better and gives the accurate prediction**.



Ada Boosting Classifier stands out as the best model for the following reasons:

1) The Ada Boosting achieved a test accuracy of 82.40%, which is the highest among all the models evaluated.

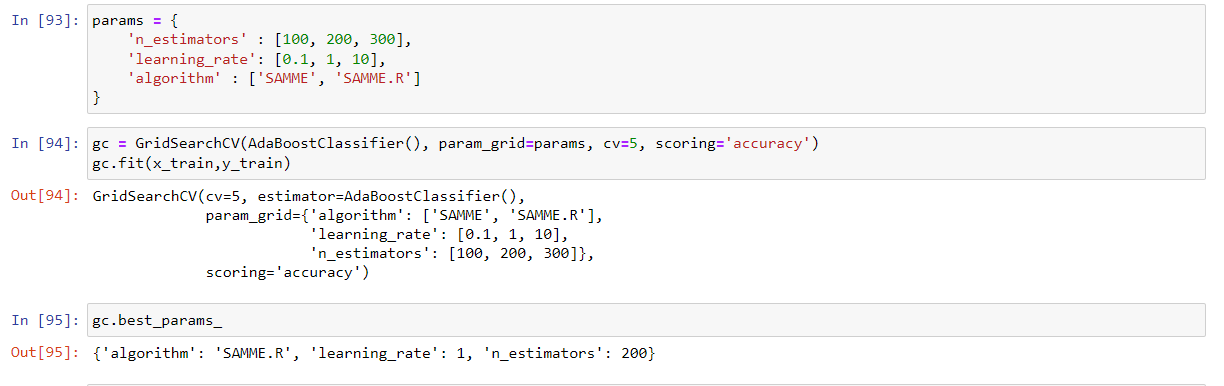
2) Strong Classification Metrics: Precision and Recall: It has high precision (0.74 for class 1.0) and recall (0.78 for class 1.0), indicating that it performs well in identifying positive cases with minimal false positives and false negatives. F1-Score: The F1-score for class 1.0 is 0.76, demonstrating a good balance between precision and recall.

3) The average cross-validation score of 0.855 is also high, showing that the model generalizes well across different subsets of the data.

4) Although the training accuracy is 83.48%, the test accuracy remains high at 82.40%, which indicates that the model is not overly complex for the data at hand, minimizing the risk of overfitting.

Hyperparameter Tuning :

Based on the cross validation score we did the model hyperparameter tuning for balance the bias and variance of the selected best model.

****

**6. Concluding Remarks :**

The Titanic dataset offers an excellent opportunity to practice binary classification and gain insights into feature engineering and model building. Through Exploratory Data Analysis, we identified key features such as **Sex**, **Pclass**, **Age**, and **Fare** as influential predictors. The pre-processing pipeline ensured that the data was clean and ready for modeling.

Among the various machine learning models, ensemble methods like **Random Forest** or **XGBoost** often perform the best for this type of problem due to their ability to handle feature interactions and reduce overfitting. However, simpler models like **Logistic Regression** can also yield strong results with proper feature selection and tuning.

By carefully following each step—from problem definition to model evaluation—we were able to predict survival outcomes effectively, demonstrating how machine learning can provide valuable insights into historical datasets.